Cell Broadband Engine™ Update

ScicomP13, Garching 19-Jul-2007

Michael Hennecke
HPC Systems Architect

Acknowledgements

- Material based on presentations by
  H. Peter Hofstee (IBM Cell/B.E. Chief Scientist)
  - With contributions by:
    - Michael Paolini
    - Brian Flachs
    - Bill Holland
    - John Easton

Special Notices

This document was developed for IBM offerings in the United States as of the date of publication. IBM may not make these offerings available in other countries, and the information is subject to change without notice. Consult your local IBM business contact for information on the IBM offerings available in your area. In no event will IBM be liable for damages arising directly or indirectly from any use of the information contained in this document.

Information in this document concerning non-IBM products was obtained from the suppliers of these products or other public sources. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. Send license inquires, in writing, to IBM Director of Licensing, IBM Corporation, New Castle Drive, Armonk, NY 10504-1785 USA.

All statements regarding IBM future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

The information contained in this document has not been submitted to any formal IBM test and is provided "AS IS" with no warranties or guarantees either expressed or implied.

All examples cited or described in this document are presented as illustrations of the manner in which some IBM products can be used and the results that may be achieved. Actual environmental costs and performance characteristics will vary depending on individual client configurations and conditions.

IBM Global Financing offerings are provided through IBM Credit Corporation in the United States and other IBM subsidiaries and divisions worldwide to qualified commercial and government clients. Rates are based on a client's credit rating, financing terms, offering type, equipment type and options, and may vary by country. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice.

IBM is not responsible for printing errors in this document that result in pricing or information inaccuracies.

All prices shown are IBM's United States suggested list prices and are subject to change without notice; reseller prices may vary.

IBM hardware products are manufactured from new parts, or new and serviceable used parts. Regardless, our warranty terms apply.

Many of the features described in this document are operating system dependent and may not be available on Linux. For more information, please check: http://www.ibm.com/systems/p/software/whitepapers/linux_overview.html

Any performance data contained in this document was determined in a controlled environment. Actual results may vary significantly and are dependent on many factors including system hardware configuration and software design and configuration. Some measurements quoted in this document may have been made on development-level systems. There is no guarantee these measurements will be the same on generally-available systems. Some measurements quoted in this document may have been estimated through extrapolation. Users of this document should verify the applicable data for their specific environment.

Special Notices (Cont.) -- Trademarks

The following terms are trademarks of International Business Machines Corporation in the United States and/or other countries: alphaWorks, BladeCenter, Blue Gene, ClusterProven, developerWorks, e business(logo), e(logo)business, e(logo)server, IBM, IBM(logo), ibm.com, IBM Business Partner (logo), IntelliStation, MediaStreamer, Micro Channel, NUMA-Q, PartnerWorld, PowerPC, PowerPC(logo), pSeries, TotalStorage, xSeries; Advanced Micro-Partitioning, eServer, Micro-Partitioning, NUMACenter, On Demand Business logo, OpenPower, POWER, Power Architecture, Power Everywhere, Power Family, Power PC, PowerPC Architecture, POWER5, POWER5+, POWER6, POWER6+, Redbooks, System p, System p5, System Storage, VideoCharger, Virtualization Engine.
A full list of U.S. trademarks owned by IBM may be found at: http://www.ibm.com/legal/copytrade.shtml.

Rambus is a registered trademark of Rambus, Inc.
XDR and FlexIO are trademarks of Rambus, Inc.
UNIX is a registered trademark in the United States, other countries or both.
Linux is a trademark of Linus Torvalds in the United States, other countries or both.
Fedora is a trademark of Redhat, Inc.
Microsoft, Windows, Windows NT and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries or both.
Intel, Intel Xeon, Itanium and Pentium are trademarks or registered trademarks of Intel Corporation in the United States and/or other countries.
AMD Opteron is a trademark of Advanced Micro Devices, Inc.
Java and all Java-based trademarks and logos are trademarks of Sun Microsystems, Inc. in the United States and/or other countries.
TPC-C and TPC-H are trademarks of the Transaction Performance Processing Council (TPPC).
SPECint, SPECfp, SPECjbb, SPECweb, SPECjAppServer, SPEC OMP, SPECviewperf, SPECapc, SPEChpc, SPECjvm, SPECmail, SPECimap and SPECsfs are trademarks of the Standard Performance Evaluation Corp (SPEC).
AltiVec is a trademark of Freescale Semiconductor, Inc.
PCI-X and PCI Express are registered trademarks of PCI SIG.
InfiniBand™ is a trademark the InfiniBand® Trade Association
Other company, product and service names may be trademarks or service marks of others.

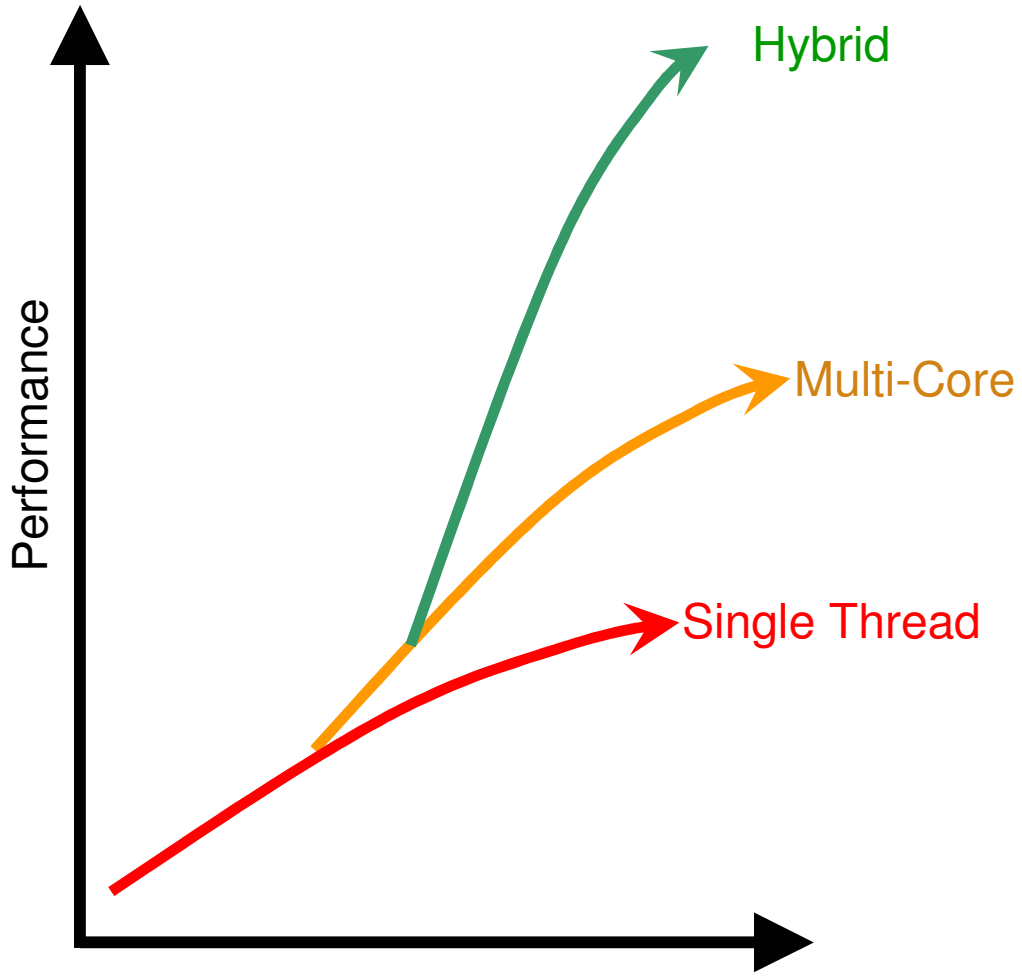
Content

- Background on Cell/B.E.
- Current IBM QS20 blade and QS2x roadmap
- LANL RoadRunner update
- Programmability

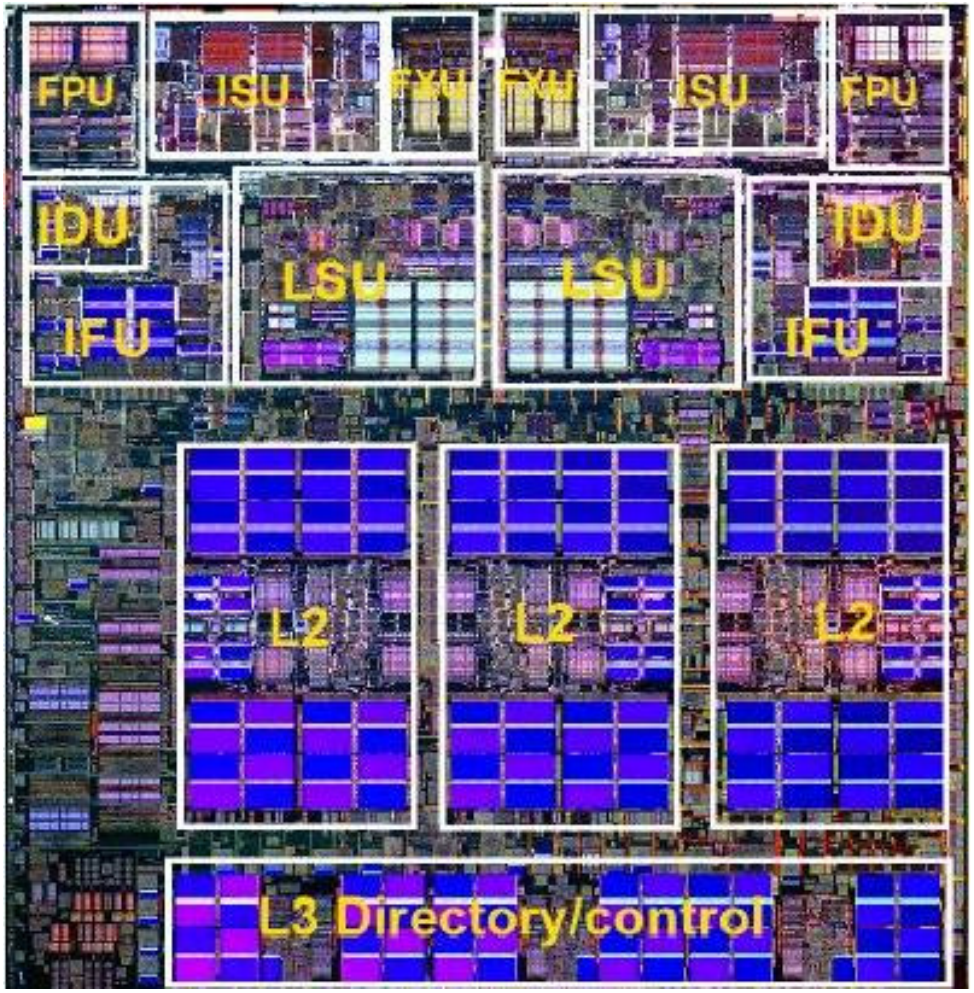
Some Background on Cell/B.E.

Microprocessor Trends

- Single Thread performance power limited
- Multi-core throughput performance extended
- Hybrid extends performance and efficiency

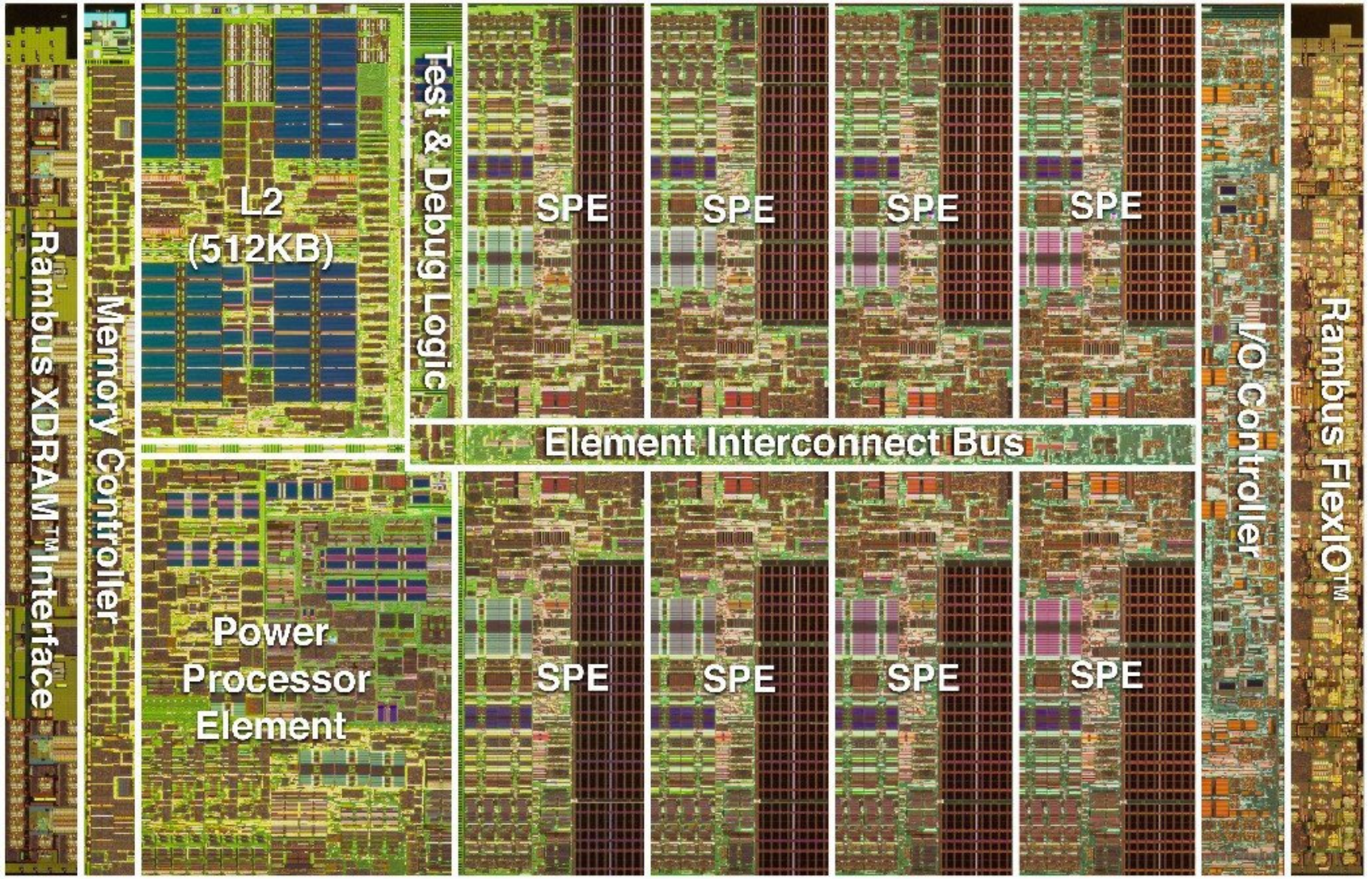


Traditional General Purpose Multi-Core Processor



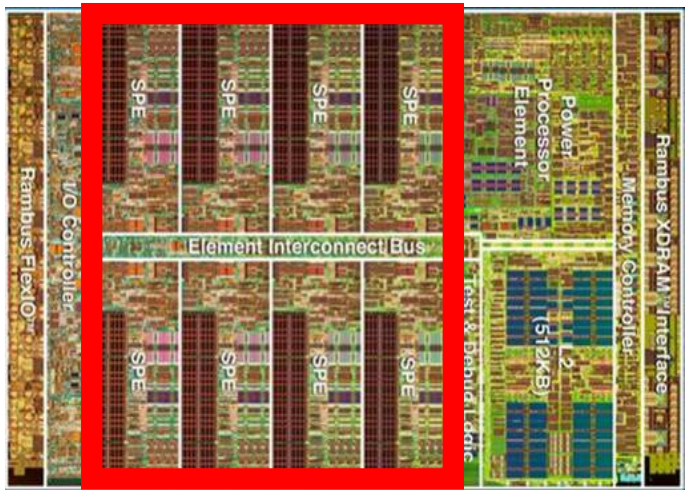
IBM Power5+

Cell Broadband Engine™: A Heterogeneous Multi-Core Architecture

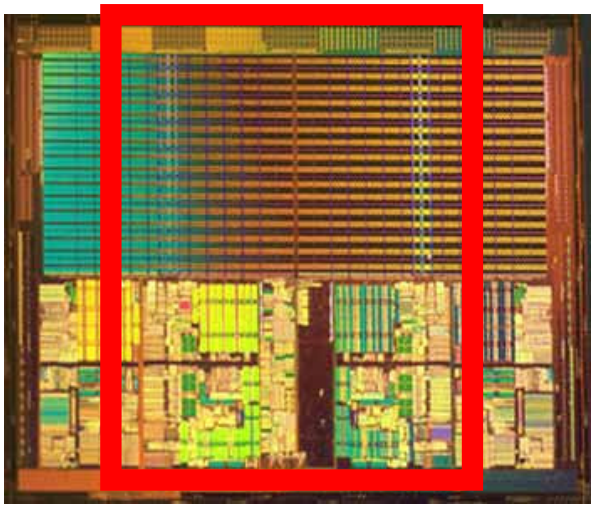


Hybrid Processor vs. Traditional General Purpose Processor Area

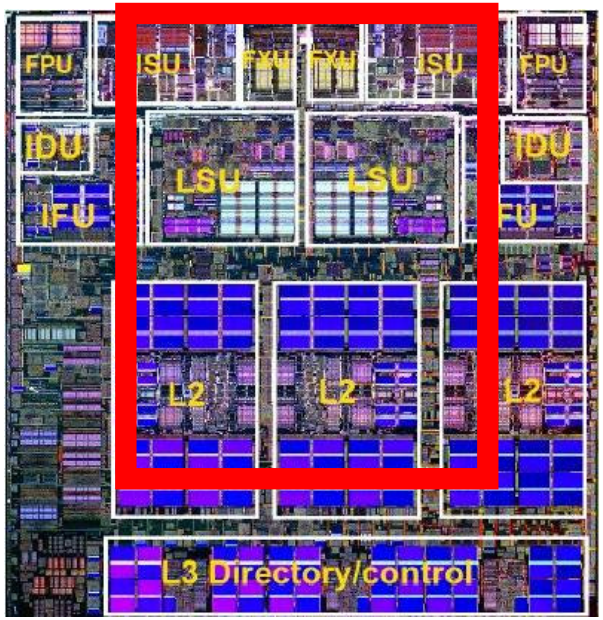
Cell BE



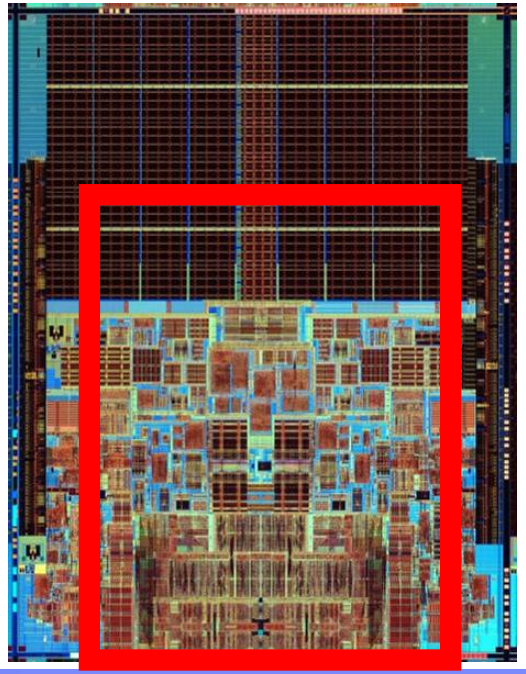
AMD



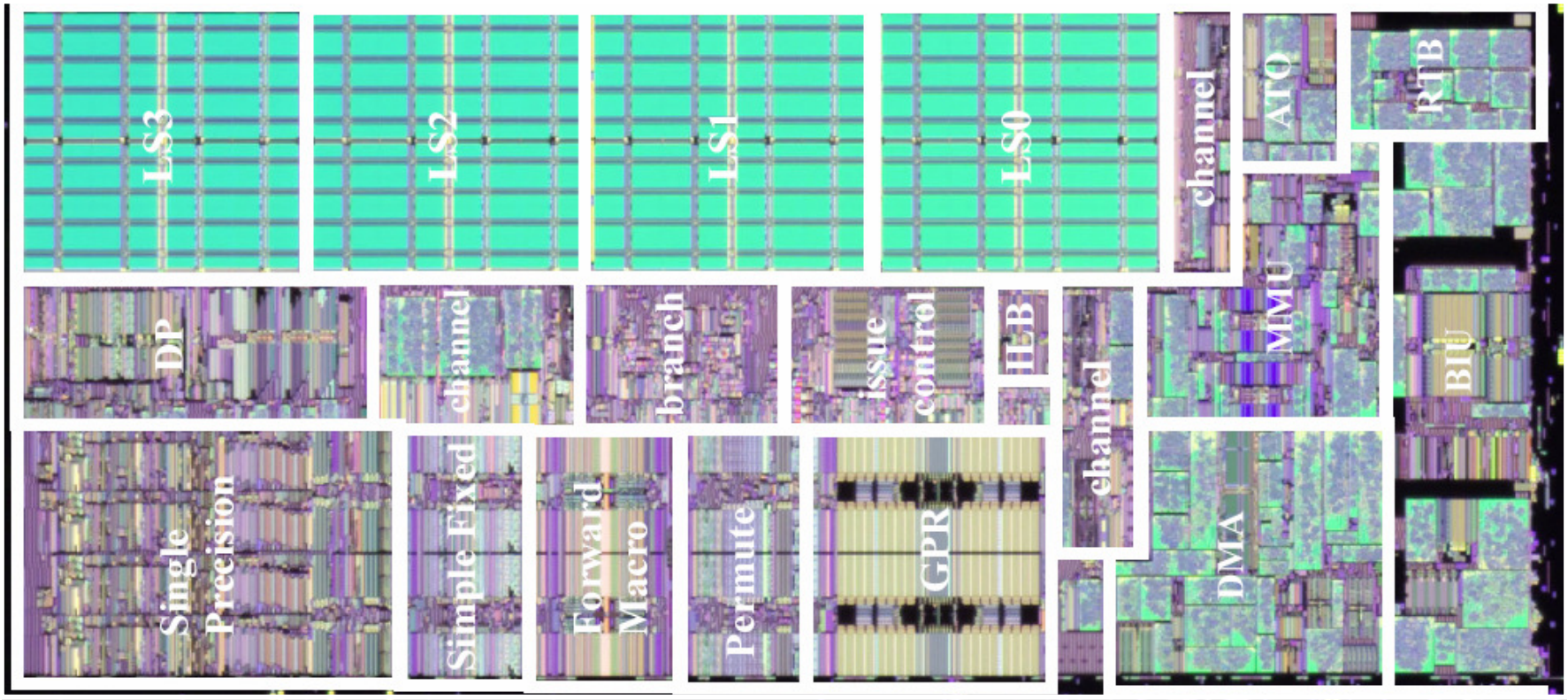
IBM



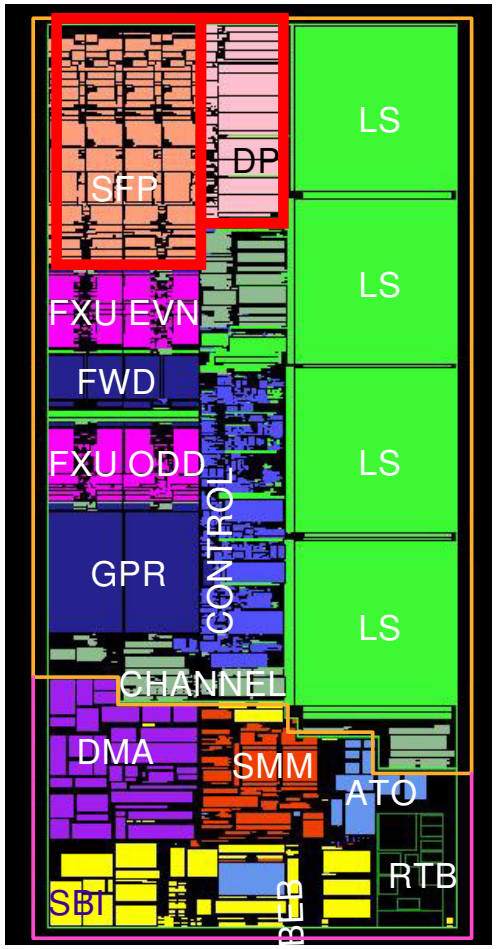
Intel



Synergistic Processor Element (SPE):



SPE Highlights



14.5mm² (90nm SOI)

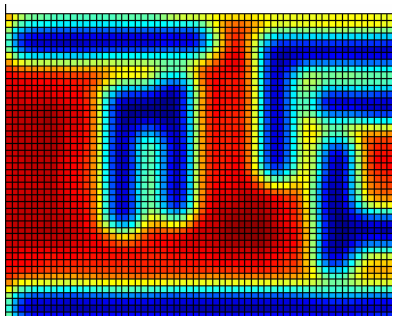
- User-mode architecture
  - No need to run the O/S
  - No translation/protection within SPU
  - DMA is full Power Architecture protect/translate
- Not just a coprocessor, has its own PC
  - RISC like organization
  - 32 bit fixed width instructions
  - Dual Issue, 11-FO4 design
  - Broad set of operations (8/16/32/64)
  - VMX-like SIMD dataflow
  - Graphics SP-Float, IEEE DP-Float
- Large unified register file
  - 128 entry x 128 bit (I&FP)
  - Deep unrolling to cover unit latencies
- 256 kB Local Store
- Flexible DMA Engine
  - Improve effective memory bandwidth
    - improving latency tolerance
    - Improve utilization of moved data
  - Vector Load/Store with Scatter/Gather

SPE Local Store

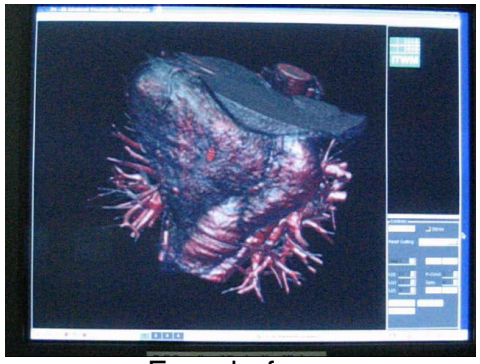
- Never misses
  - Both instruction fetch and data
  - No tags, backing store, or prefetch engine
  - Predictable real-time behavior
  - Less wasted bandwidth
  - Easier programming models to achieve very high performance
    - Level of software controlled memory hierarchy
      - Local Memory
      - Stream buffers
      - Vector register file
    - Software managed caching
      - Different data types can have different:
        - Caches – no collisions
        - Replacement policies, Line sizes, Associativities
      - Much higher hit rates than with hardware caches
  - DMA's are fast to setup
    - almost like normal load instructions
    - Can move data from one local store to another
- No translation from SPU ports
  - Multiuser operating system is running on control processor
  - Can be mapped as system memory - cached copies are non-coherent wrt SPU loads/stores

Cell BE based Systems: SCEI, IBM, Mercury, …

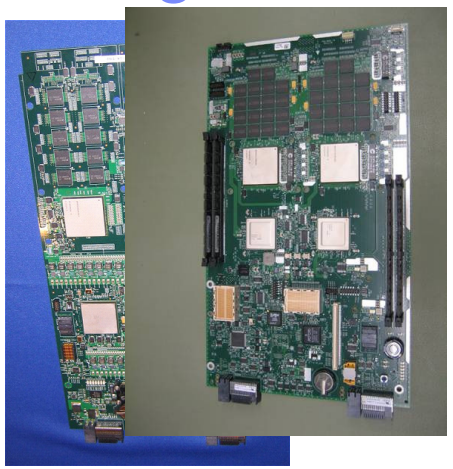
Many Applications for Cell/B.E. Beyond Gaming

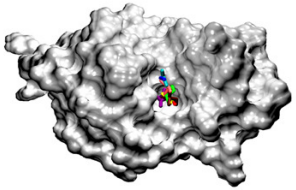


Mercury/Mentor Graphics
45nm OPC tool



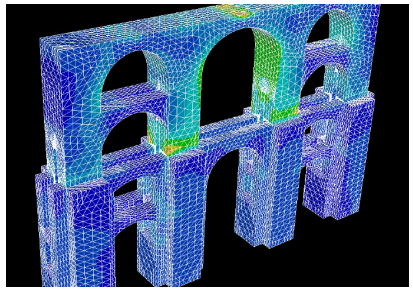
Fraunhofer
PV4D Medical Imaging



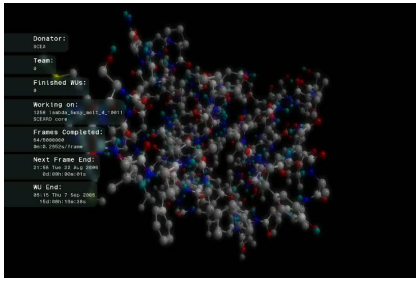


Boston Univ.
Bioinformatics: FBDD

I.B.M. to Build Supercomputer Powered by Video Game Chips
By JOHN MARKOFF
(NY Times): September 7, 2006



Structural Analysis
digitalmedics.de



SCEI / Pande (Stanford)
folding@home PS3 client

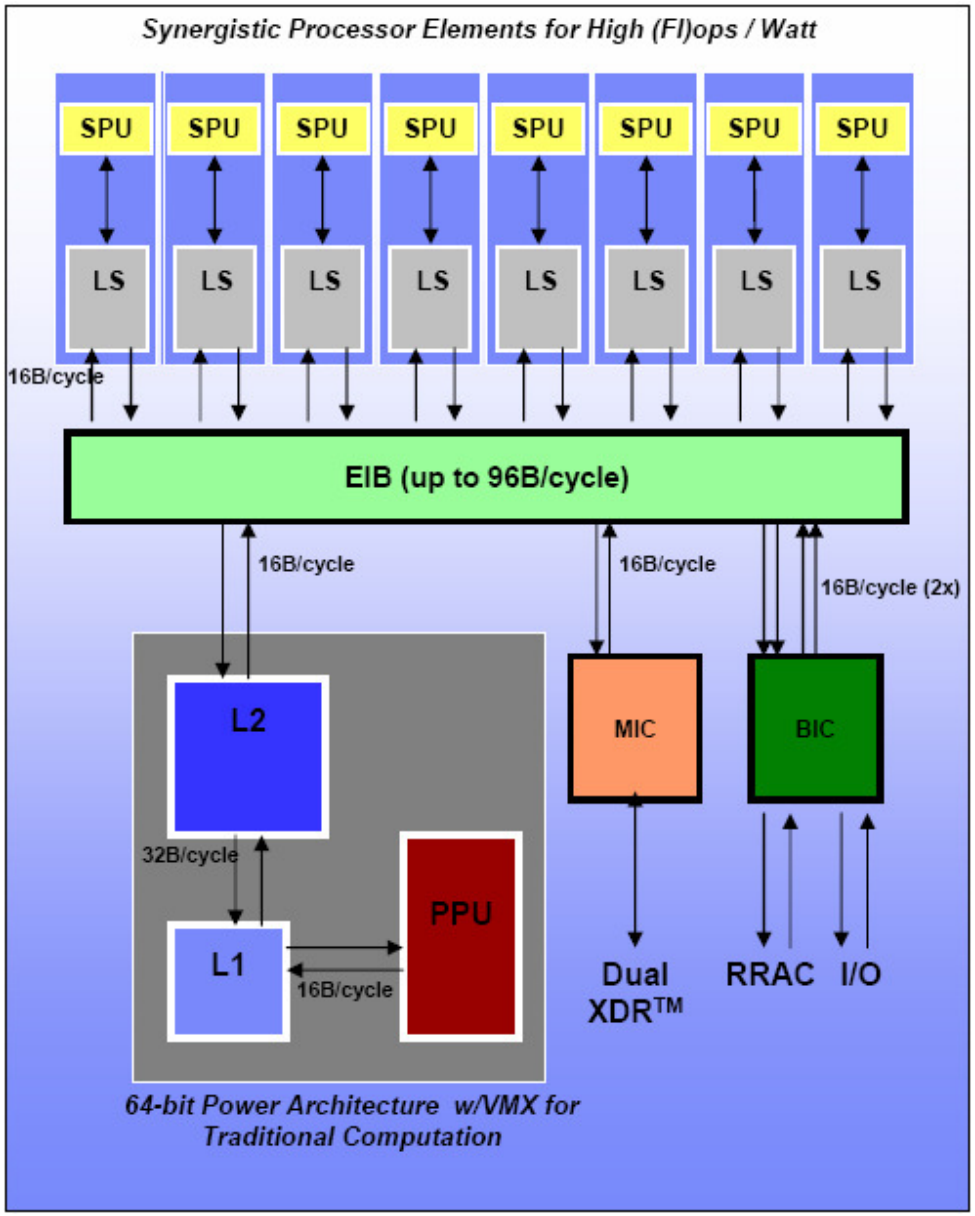
Rapidmind(TM) / RTT

IBM iRT raytracer prototype

Current IBM QS20 blade and QS2x roadmap

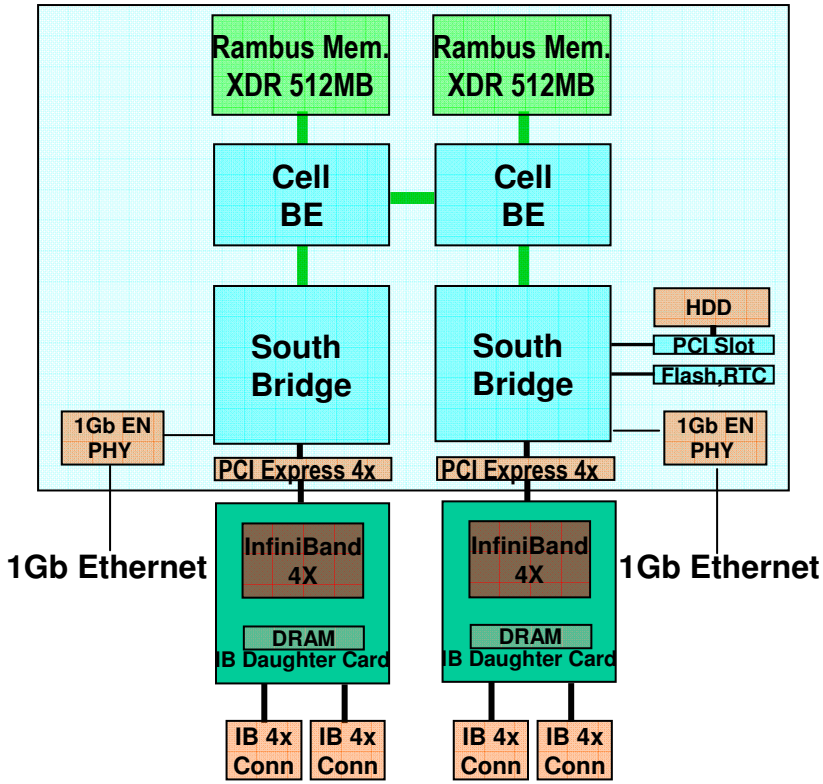
Cell processor overview

- One Power-based PPE, with VMX
  - 32/32kB I/D L1, and 512kB L2
  - dual issue, in order PPU, 2 HW threads
- Eight SPEs, with up to 16x SIMD
  - dual issue, in order SPU
  - 128 registers (128b wide)
  - 256 kB local store (LS)
  - 2x 16B/cycle DMA, 16 outstanding req.
- Entity Interconnect Bus (EIB)
  - 4 rings, 16B wide at 1:2 clock
  - 96B/cycle peak, 16B/cycle to memory
  - 2x 16B/cycle BIF and I/O
- External communication
  - Dual XDR™ memory controller (MIC)
  - Two configurable bus interfaces (BIC)
    - Classical I/O interface
    - SMP coherent interface



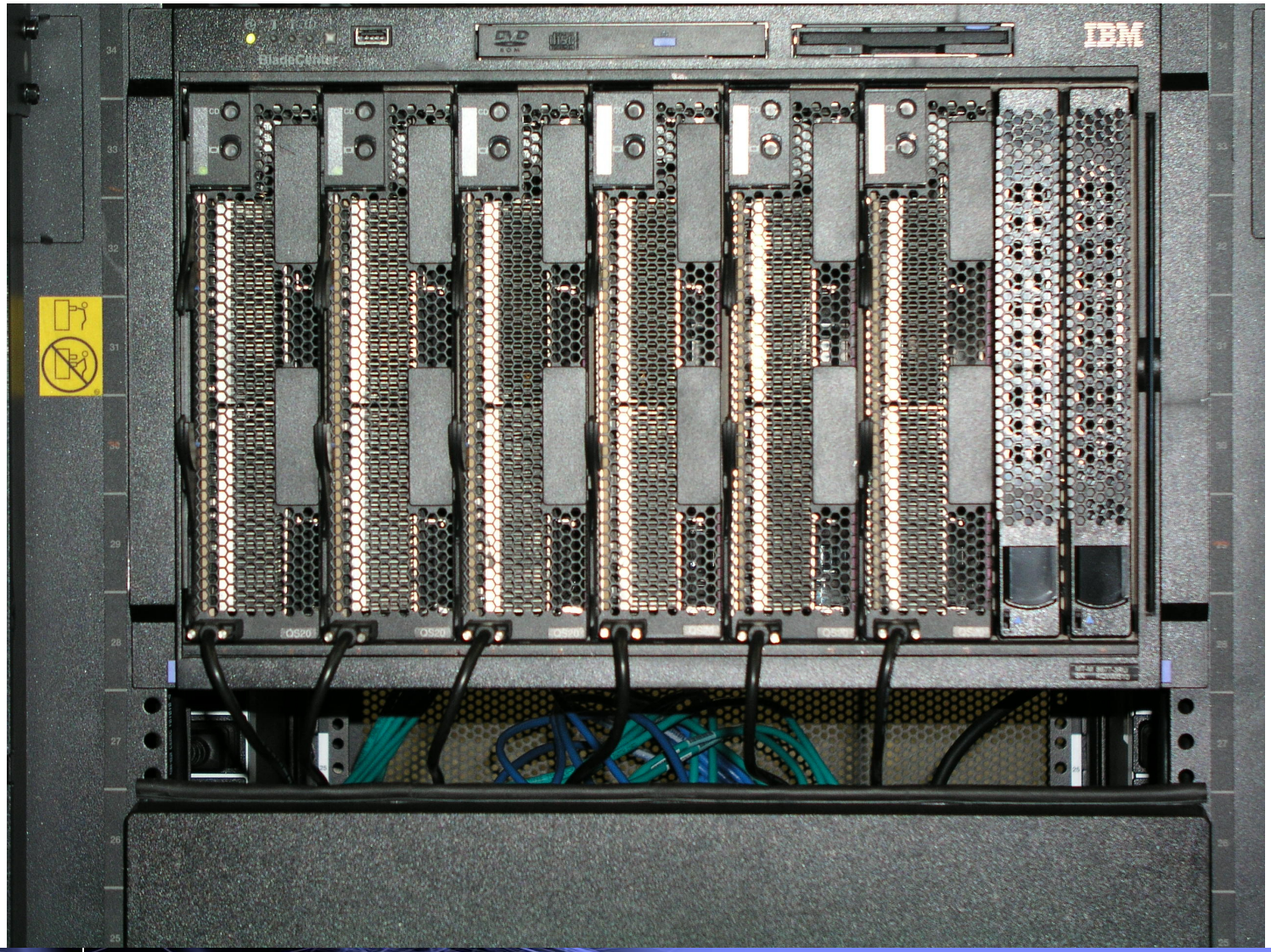
IBM Bladecenter QS20 blade – T/M 0200/AC1

- QS20 Cell/B.E. blade @ 3.2 GHz
  - Double-wide blade (max 7 per BC-1 chassis)
  - Two Cell/B.E. "sockets", 90nm SOI
    - each has 1 PPE and 8 SPEs
  - 1GB Rambus XDR memory (2x 512 MB)
  - One 40GB IDE hard disk
  - Two 1GigE ports into BC-1 backplane
- QS20 f/c 2945: InfiniBand card kit
  - Dual-port IB 4x SDR daughter card
    - attaches to PCI-e 4x connector
  - Up to two cards per QS20 blade
  - Needs external InfiniBand switch



- Peak performance (SPEs only):
  - 128bit=4x32bit SIMD * 2 (FMA) * 8 SPUs * 3.2 GHz = 204.8 GFlop/s per *socket*
  - 2 sockets per blade = 409.6 GFlop/s per *blade*
  - 7 blades per BC-1 chassis = 2.9 TFlop/s per *chassis* (all numbers for 32bit floating point)

Front view of a BC-1 chassis with six QS20 blades



Cell Broadband Engine Architecture Blades

IBM BladeCenter QS20 and beyond

GA September 2006

BladeCenter QS20
- 2 Cell/B.E. processors
- 1 PPE + 8 SPE
- SP: 460 GFLOPS per blade
- DP: 42 GFLOPS per blade
- 1 GB memory

SDK 1.1

Available July 2006

Target availability: 4Q2007

BladeCenter QS2X
- 2 Cell/B.E. processors
- 1 PPE + 8 SPE
- SP: 460 GFLOPS per blade
- DP: 42 GFLOPS per blade
- Next Generation I/O chip
- 2 GB memory

SDK 2.1

Available: March 2007

SDK 3.0

Target release: September 2007

Target availability: 1H2008

BladeCenter QS2Y
- 2 CBEA-compliant processors
- 1 PPE + 8 eDP SPE
- SP: 460 GFLOPS per blade
- eDP: 217 GFLOPS per blade
- Up to 32 GB memory
- PCI Express™ x16 slots

SDK 4.0

Target release: March 2008

Target availability: 1H2010

BladeCenter QS2Z
- First CBEA teraflop processor
- 2 PPE' + 32 eSPE
- Power Architecture compliant
- ~2 TFLOPS SP per blade
- ~1 TFLOPS DP per blade
- Next generation memory technology

SDK 5.0

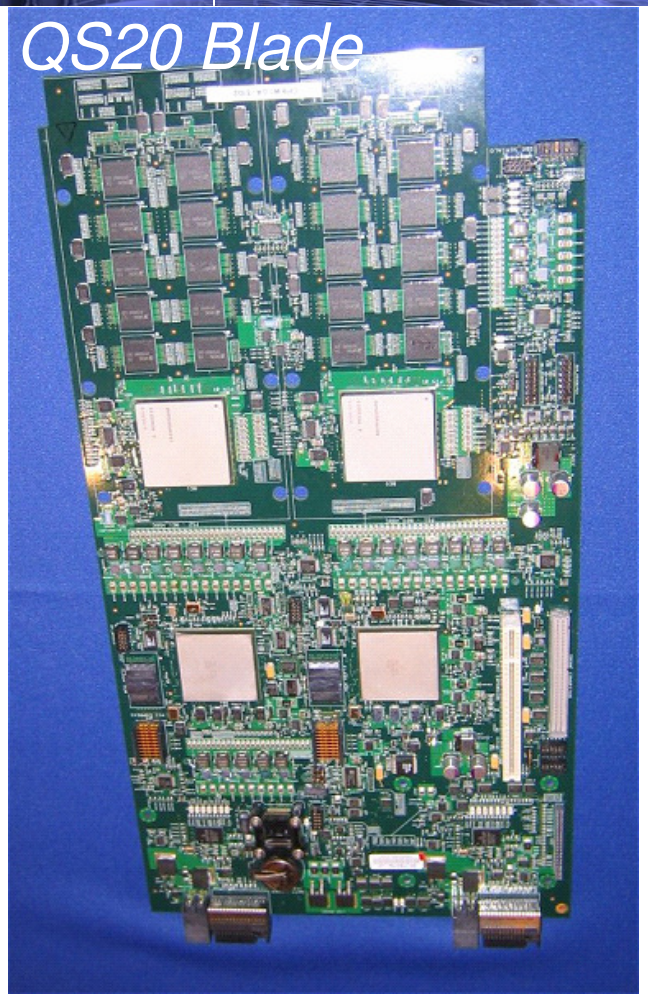
Target release: December 2008

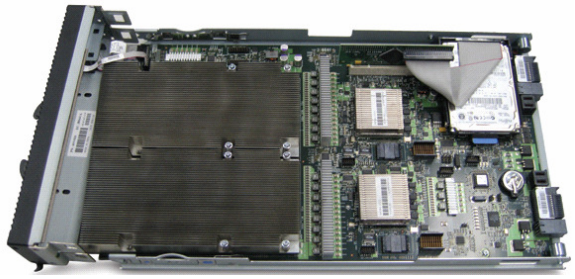
Committed
Concept

2006 **2007** **2008** **2009-2010**

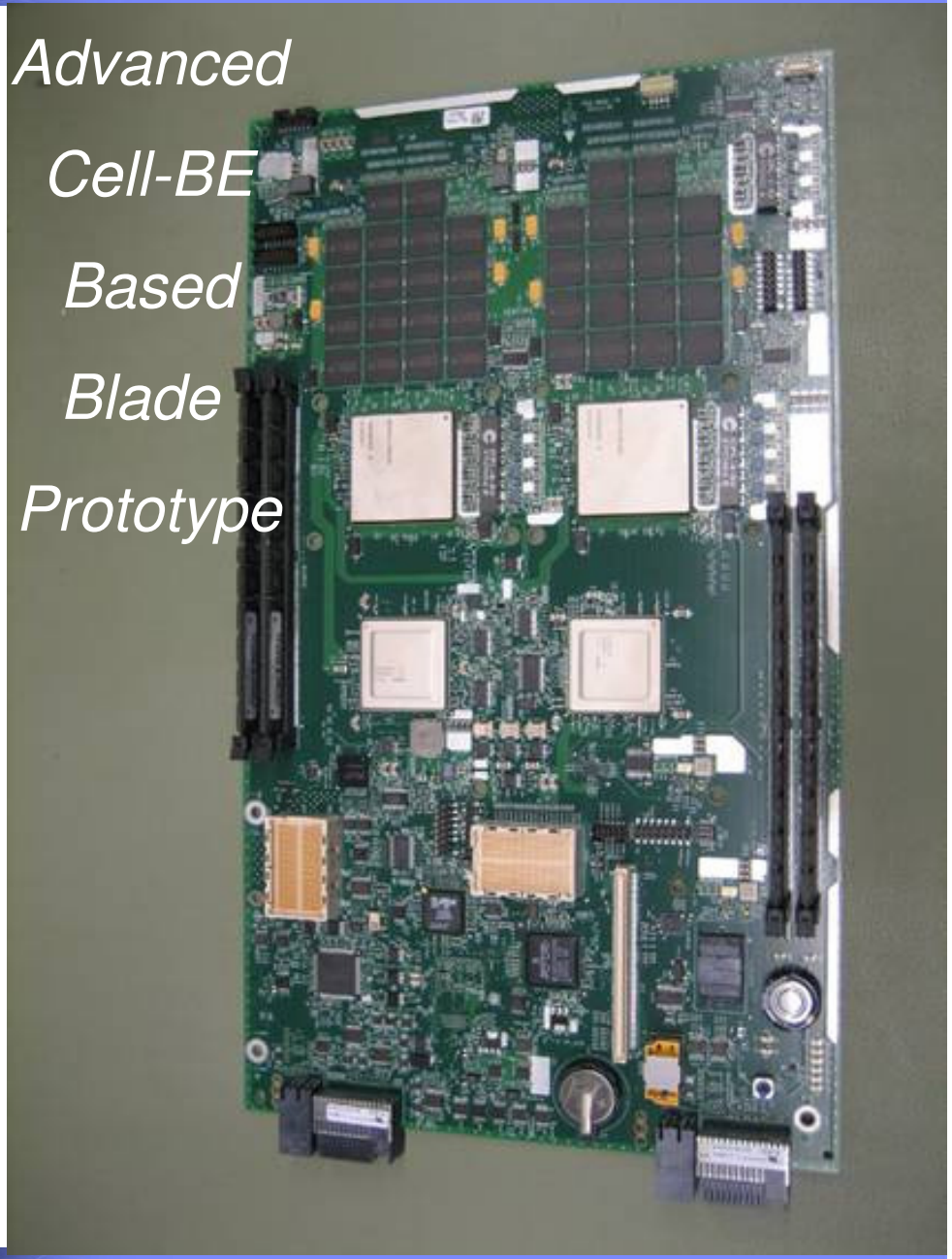
All future dates and specifications are estimations only; Subject to change without notice.

QS20 Blade





Advanced Cell-BE Based Blade Prototype



Software Development Kit (SDK) 2.1 – March 2007

- SDK 2.1 Installation Guide
- SDK 2.1 Programmer's Guide
- Cell Broadband Engine Programming Tutorial

- IBM Full-System Simulator

- XL C/C++ compiler
- SIMD math, MASS, and MASS/V libraries
- Sample libraries and files; tools and utilities
- Eclipse IDE for Cell Broadband Engine

- Can be downloaded from IBM developerWorks…
  - http://www-128.ibm.com/developerworks/power/cell/

Software Development Kit (SDK) roadmap

SDK 3.0
- First production-grade release
- Support for Linux Enterprise Distribution
- Enhanced programmer Productivity

Target release: September 2007

SDK 4.0
- Enhanced blade-to-blade collaboration framework
- Fortran 11.1 GA
- Broader Ecosystem Support

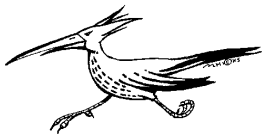
Target release: March 2008

- Product Quality/Performance
- XL C/C++ GA v9.0 (Dual Source)
- XL Fortran 11.1 (Beta)
- Support for Linux Enterprise Distro
- Programming Model ALF/DaCS GA
- Tech Preview of Hybrid Programming Model GA (ALF/DaCS)
- Base library enhancements
- eDP & Industry Libraries
- IDE - Eclipse
- Performance Analysis Tools
- CBEA System Simulator enhancements
- Security implementation

- XL C/C++ GA v9.0
- XL Fortran 11.1 GA
- Additional Linux Distro Support
- Major GNU Toolchain updates
- Enhanced blade-to-blade collaboration framework GA (ALF/DaCS)
- eDP & additional Industry Libraries
- Base library enhancements

All future dates and specifications are estimations only; Subject to change without notice.

LANL RoadRunner Update



Roadrunner Project Charter - Phase 3

- Demonstrate 1 PF sustained performance in 2Q08 and deliver a system for production in 3Q08
- IBM Global Engineering Solutions is responsible for Development, Manufacturing, Maintenance and Support of the RR system
  - TriBlade is not currently in the IBM Blade Center product plans
    - it may be added at a later date
  - This system is a Custom IBM Machine Type
    - Not an generally available product
- Limited Availability through IBM GES

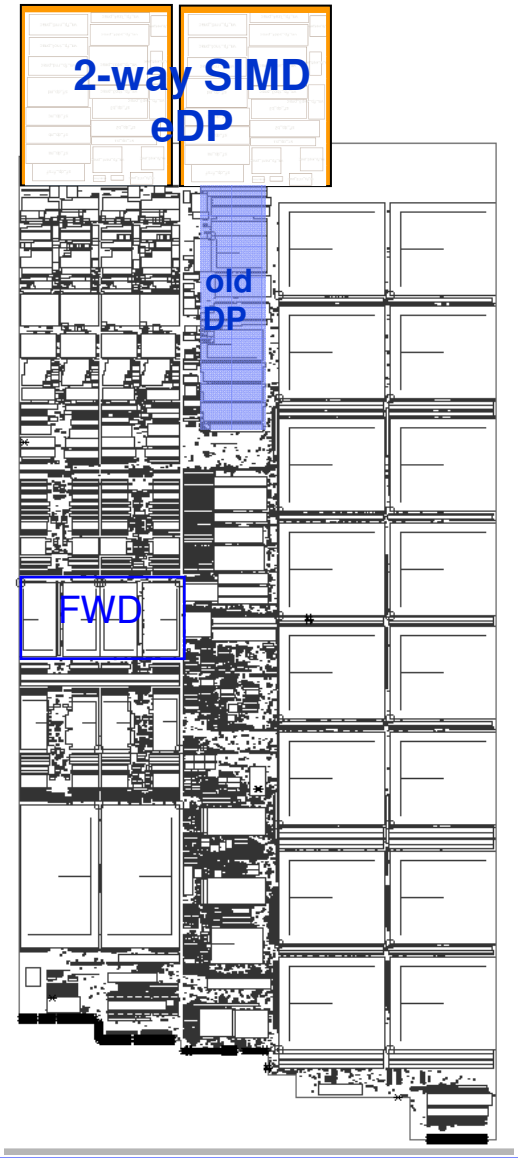
Overview: Integrated Compute Node

- 3 processor blades + 1 expansion board
  - Custom expansion card
    - Six PCIe x8 links, including HSDC and PCIe x8 LP slot
  - Dual PCIe x8 flex cables to each I/O Hub in QS22
  - Custom and existing mechanicals for 4 slot assembly
- Use of LS21 and QS22 card assemblies
  - As-is board and components
  - Modified firmware
  - Modified mechanicals
- 3 Compute Nodes will fit into 1 BC-H chassis
- Entire Compute Node will be Field Replaceable

Enhanced BE – An HPC Cell Implementation

now featuring DDR2 and an enhanced SPE

Challenge:	***Response:***
2GB Memory Limit	DDR2 allows upto 16 GB
Still want 25 GB/s	Upto DDR2-800 Many more pins
25.6 Gflops DP/BE ▪13 Cycle DP Latency ▪6 Cycle Stall ▪No Dual Issue w/DP	102 Gflops DP/BE ▪9 Cycle DP Latency ▪Fully Pipelined DP ▪Dual Issue w/DP
IEEE compliance ▪Denormal Inputs -> 0 ▪Default NaNs	IEEE compliance is improved ▪Denormal Support ▪Expected NaNs
Up to 10% perf. loss in DP Compare Emulation	5 new DP compare instructions – SPU ISA v1.2



Cell eDP / AMD (Dual Core Opteron, IB-DDR)

AMD Host Blade + Expansion

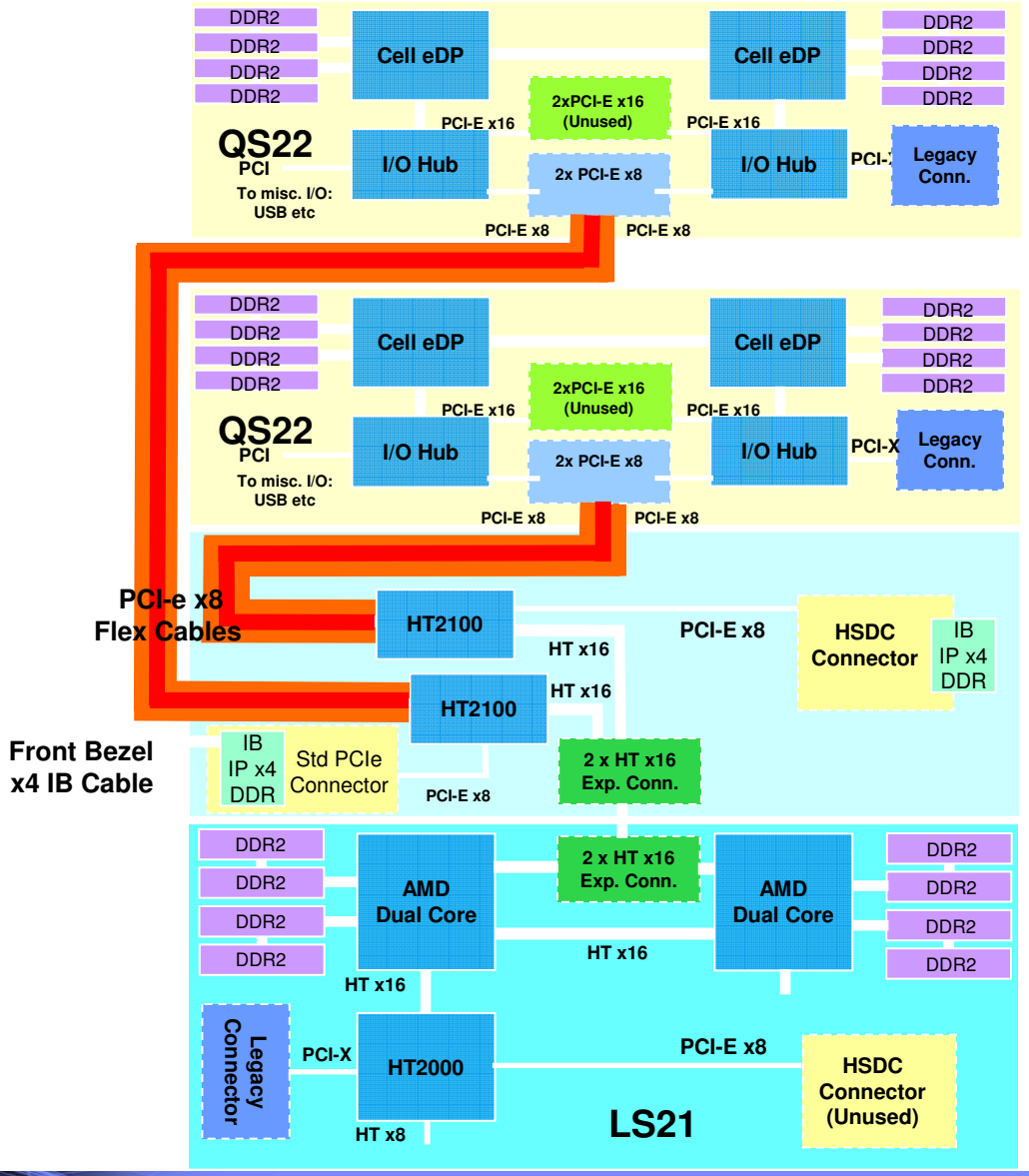
- Dual socket dual core AMD Opteron 1.8 GHz (2 x 7.2 GFlops)
  - LS21 + 2 by HT 16x connector
- DDR2 direct attach DIMM channels
  - 8GB
  - 10.7 GB/s/socket (0.48 B/Flop)
- **New Expansion Card**
  - 2 HT2100 HT<->PCI-e bridges

QS22 Accelerator Blade

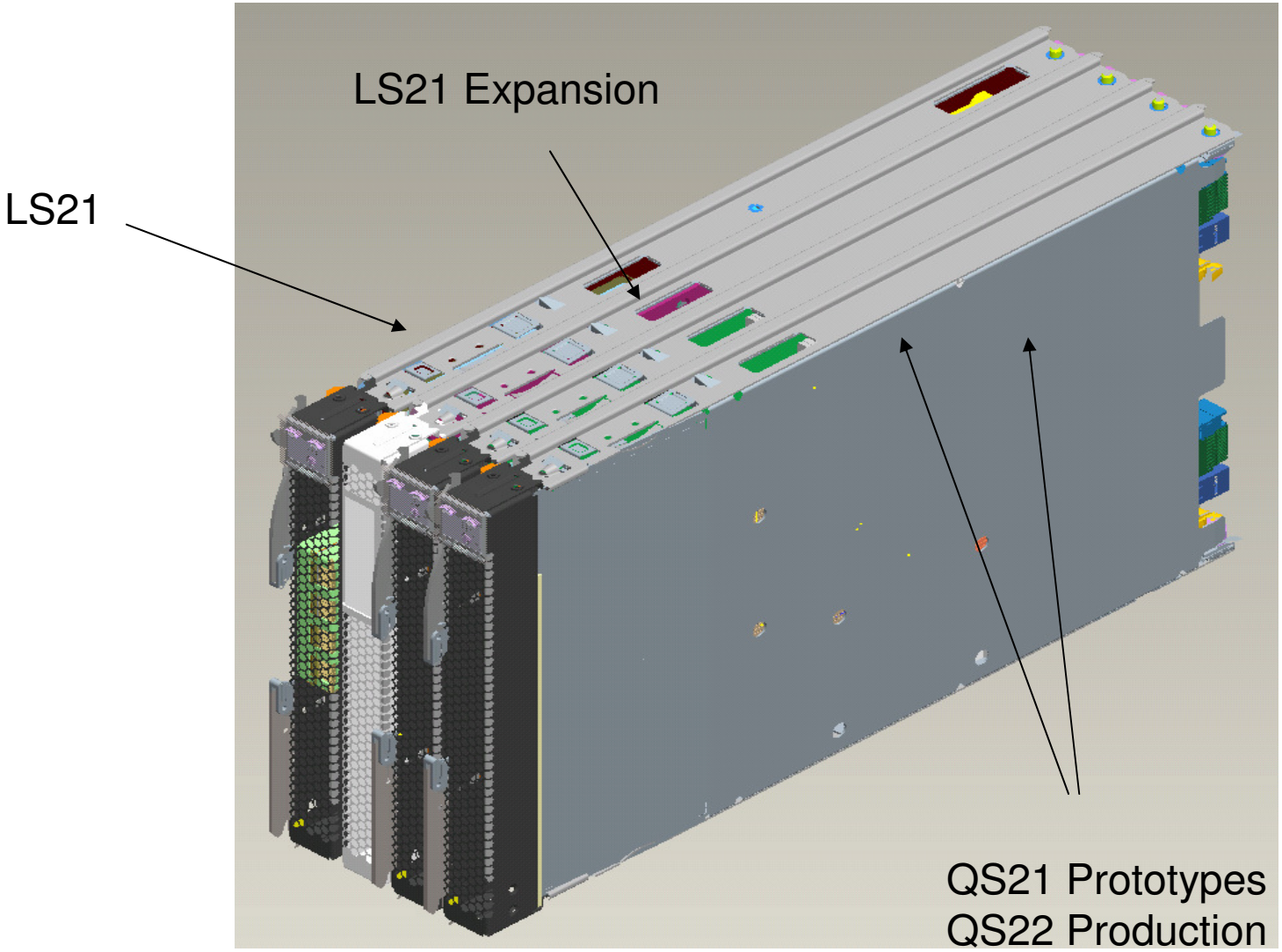
- Dual Cell eDP Sockets
  - 204.8 GFlop/s @ 3.2Ghz (2 x 102.4 GFlop/s)
- DDR2 direct attach DIMM channels
  - 8 GB (4GB/Cell eDP)
  - 25.6 GB/s per Cell eDP chip (0.25 B/Flop)

AMD Host to Cell eDP connectivity

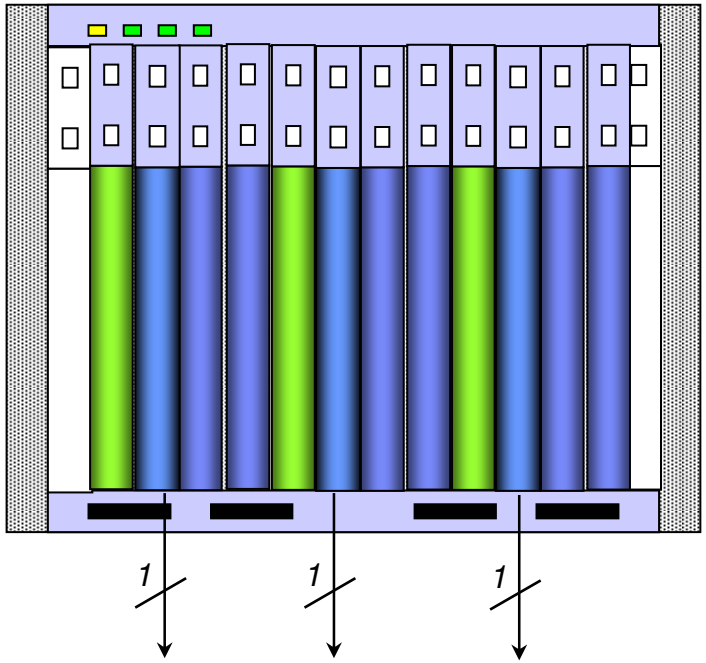
- Two x8 PCIe Host to CB2 links
- ~2+2 GB/s/link → ~4+ 4 GB/s total POR



Full Package



Blade Center Chassis

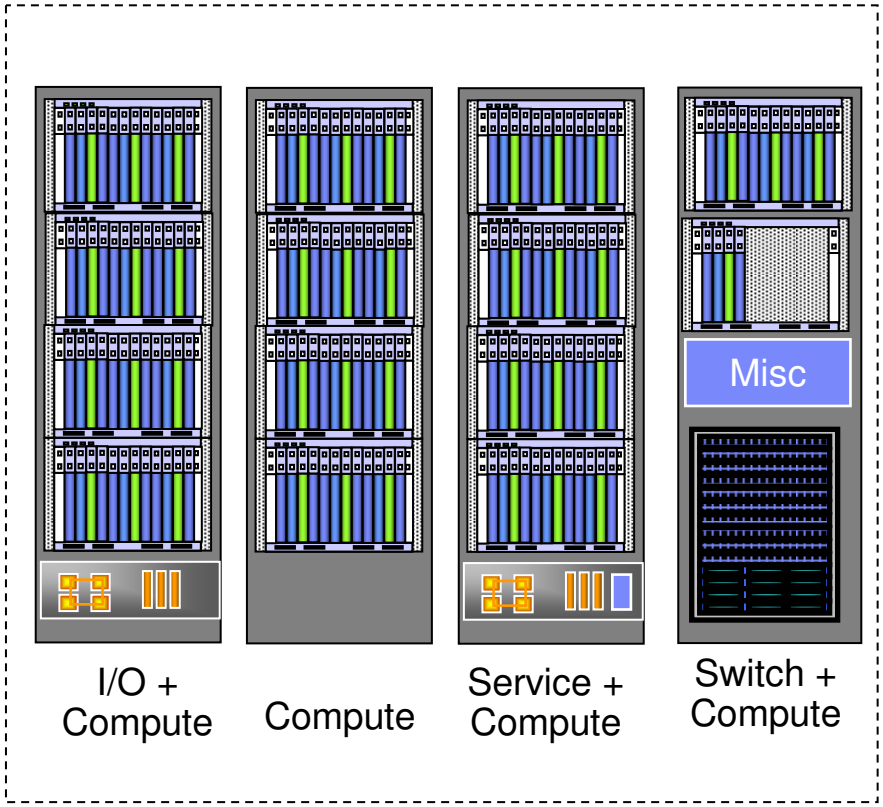


IB 4x DDR

1st stage switches

- BC-H
  - 3 Tri-Blades (4 slots with expansion card)
  - 3 OS Images = 3 hybrid nodes (HN)
- Capability
  - Host: 43.2 GF
  - Cell: 1.23 TF
- Connectivity
  - 3 HNs x 1 planes x (2 + 2) GB/s
  - 3 IB 4x DDR copper cables
  - 12 GB/s bidir. peak bandwidth

Connected Unit – Dense CU



- CU Capability
  - Host: 2.6 TF
  - Cell: 76.6 TF
  - IO: 25.6 GB/s

- CU size driven by 1st Stage Switch
  - Using 288 port Voltaire ISR9288
    - IB 4x DDR
  - Half Bisection to other CU's
    - 192 ports Down, 96 ports Up
- 62 BC-H,
  184 Hybrid Compute Nodes
  - 184 LS21
  - 368 CB2
- 6 x3755 IO nodes
- 1 Service Node
  - Out of band
- 16 Racks
- 4 Rack Types
  - Rack integrated unit

Programmability

Fundamentals of Programmability

- It is always about ROI … how much **performance** for how much **effort**. Many factors go into this.
  - Application-level success is final arbiter.
- Full architecture/implementation disclosure is important, a large enough ecosystem is critical.
  - Architecture should mean that a program not only runs from one implementation to another, but runs well.
- Predictability is crucial. Nothing slows tuning more than having performance change (degrade) for mysterious reasons.
  - Even with full disclosure, complexity can be a major hindrance.
- Memory management is usually the biggest factor. Throughout the hierarchy:
  - Coherence
  - Size
  - Bandwidth
  - Minimum efficient transfer granule
  - Alignment granule

  have a strong effect on the amount of programming effort required.

Programming Models for single-node Cell/B.E.
(only considering Cell-Unique aspects e.g. not SIMD)

- Pure Streaming
  - Data flows from one SPE to another
  - Used initially, not used much any more (load balancing)
- Pure function offload model
  - SPEs accelerate PPE threads
  - Not so effective, too much load on PPE
- Task queue model
  - Dominant now, several variants (e.g. OpenMP, Charm++)
  - Some versions use multiple queues
- (Quasi-)Sequential model
  - Various forms of compiler based auto-parallelization
  - Runtime-based parallelization … e.g. Rapidmind
- Highly threaded models (multiple threads per SPE)
  - E.g. SPURS

Each of these execution models allows a variety of languages and compiler approaches.

Some Cell/B.E. references

- Cell Broadband Engine documentation @ IBM developerWorks:
  - http://www-128.ibm.com/developerworks/power/cell/
- Cell Broadband Engine technology @ IBM alphaWorks:
  - http://www.alphaworks.ibm.com/topics/cell
- Cell at IBM Research:
  - http://www.research.ibm.com/cell/
- Cell at IBM Microelectronics:
  - http://www-306.ibm.com/chips/techlib/techlib.nsf/products/Cell

  (also has the *ISSCC* and *Microprocessor Report* articles on Cell)
- Linux on Cell at Barcelona Supercomputing Center:
  - http://www.bsc.es/projects/deepcomputing/linuxoncell/
- Cell presentation at Sony Research:
  - http://www.research.scea.com/research/html/CellGDC05/
- IBM Systems Journal – Online Game Technology
  - http://www.research.ibm.com/journal/sj

Questions ?